arXiv:1304.6164v1 [math.PR] 23 Apr 2013

A note on the CLT of the LSS for sample covariance matrix from a spiked population model

Qinwen Wang and Jack W. Silverstein and Jian-feng Yao

Qinwen Wang
Department of Mathematics
Zhejiang University
e-mail: wqw8813@gmail.com

Jack W. Silverstein
Department of Mathematics
North Carolina State University
e-mail: jack@ncsu.edu

Jianfeng Yao
Department of Statistics and Actuarial Science
The University of Hong Kong
Pokfulam, Hong Kong
e-mail: jeffyao@hku.hk

Abstract: In this note, we establish an asymptotic expansion for the centering parameter appearing in the central limit theorems for linear spectral statistic of large-dimensional sample covariance matrices when the population has a spiked covariance structure. As an application, we provide an asymptotic power function for the corrected likelihood ratio test in Bai et al. (2009) used for testing the presence of spike eigenvalues in the population covariance matrix. This result generalizes a formula provided in Onatski et al. (2011) where only one simple spike exists.

AMS 2000 subject classifications: Primary 60F05; secondary 62H15.
Keywords and phrases: Large-dimensional sample covariance matrices, Spiked population model, Central limit theorem, Centering parameter, factor models.

1. Introduction

Let (Σ_p) be a sequence of $p \times p$ non-random and nonnegative definite Hermitian matrices and let (w_{ij}), $i, j \geq 1$ be a doubly infinite array of i.i.d. complex-valued random variables satisfying

$$\mathbb{E}(w_{11}) = 0, \quad \mathbb{E}(|w_{11}|^2) = 1, \quad \mathbb{E}(|w_{11}|^4) < \infty.$$

Write $Z_n = (w_{ij})_{1 \leq i \leq p, 1 \leq j \leq n}$, the upper-left $p \times n$ block, where $p = p(n)$ is related to n such that when $n \to \infty$, $p/n \to y > 0$. Then the matrix $S_n = \frac{1}{n}\Sigma_p^{1/2} Z_n Z_n^* \Sigma_p^{1/2}$ can be considered as the sample covariance matrix of an i.i.d. sample $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of p-dimensional observation vectors $\mathbf{x}_j = \Sigma_p^{1/2}\mathbf{u}_j$ where

$\mathbf{u}_j = (w_{ij})_{1\le i\le p}$ denotes the j-th column of Z_n. Note that for any semi-positive definite Hermitian matrix A, $A^{1/2}$ denotes a Hermitian square root and we call the *spectral distribution* (SD) the distribution generated by its eigenvalues.

Assume that the SD H_n of Σ_p converges weakly to a nonrandom probability distribution H on $[0,\infty)$. It is then well-known that the SD F^{S_n} of S_n, generated by its eigenvalues $\lambda_{n,1} \ge \cdots \ge \lambda_{n,p}$, converges to a nonrandom limiting SD G (Marčenko and Pastur, 1967; Silverstein, 1995). The so-called *null case* corresponds to the situation $\Sigma_p \equiv I_p$, so $H_n \equiv \delta_1$ and the limiting SD is the seminal Marčenko-Pastur law G_y with index y and support $[a_y, b_y]$ where $a_y = (1-\sqrt{y})^2$, $b_y = (1+\sqrt{y})^2$, and an additional mass at the origin if $y > 1$ (Marčenko and Pastur (1967)).

In this paper we consider the *spiked population model* introduced in Johnstone (2001) where the eigenvalues of Σ_p are

$$\underbrace{\alpha_1,\ldots,\alpha_1}_{n_1},\ldots,\underbrace{\alpha_k,\ldots,\alpha_k}_{n_k},\underbrace{1,\ldots,1}_{p-M}. \tag{1.1}$$

Here M and the multiplicity numbers (n_k) are fixed and satisfy $n_1+\cdots+n_k = M$. In other words, all the population eigenvalues are unit except some fixed number of them (the spikes). The model can be viewed as a finite-rank perturbation of the null case. Obviously, the limiting SD G of S_n is not affected by this perturbation. However, the asymptotic behaviour of the extreme eigenvalues of S_n is significantly different from the null case. The analysis of this new behaviour of extreme eigenvalues has been an active area in the last years, see e.g. Baik et al. (2005), Baik and Silverstein (2006), Paul (2007), Bai and Yao (2008), Benaych-Georges et al. (2011), Benaych-Georges and Nadakuditi (2011) and Bai and Yao (2012). In particular, the base component of the population SD H_n in the last two references has been extended to a form more general than the simple Dirac mass δ_1 of the null case.

For statistical applications, besides the principal components analysis which is indeed the origin of spiked models (Johnstone (2001)), large-dimensional strict factor models are equivalent to a spiked population model and can be analyzed using the above-mentioned results. Related recent contributions in the area include, among others, Kritchman and Nadler (2008, 2009), Onatski (2009, 2010, 2012), Nadakuditi and Silverstein (2010) and Passemier and Yao (2012) and almost all of them concern the problem of estimation and testing the number of factors (or spikes).

In this note, we analyze the effects caused by the spike eigenvalues on the fluctuations of linear spectral statistics of the form

$$T_n(f) = \sum_{i=1}^{p} f(\lambda_{n,i}) = F^{S_n}(f)\ , \tag{1.2}$$

where f is a given function. Similarly to the convergence of the SD's, the presence of the spikes does not prevent a central limit theorem for $T_n(f)$; however as we will see, the centering term in the CLT will be modified according to the

values of the spikes. As this term has no explicit form, our main result is an asymptotic expansion presented in Section 2. Section 3 explains in detail the contour that appears in the main results. To illustrate the importance of such expansions, we present in Section 4 an application for the determination of the power function for testing the absence versus presence of spikes. The Appendix contains some technical derivations.

2. Centering parameters in the CLT of the LSS from a spiked model

Fluctuations of linear spectral statistics of form (1.2) are indeed covered by a central limit theory developed in Bai and Silverstein (2004). The theory was later improved by Pan and Zhou (2008) where the restriction $E(|w_{11}|^4) = 3$ matching the real Gaussian case was removed.

Let $f_1, \ldots, f_L$ be L functions analytic on an open domain of the complex plan including the support of the limiting SD. These central limit theorems state that the random vector

$$(X_n(f_1), \cdots, X_n(f_L)) \ ,$$

where

$$X_n(f) = p\left[F^{S_n}(f) - F^{y_n, H_n}(f)\right] = p\int f(x) d(F^{S_n} - F^{y_n, H_n})(x) \ ,$$

converges weakly to a Gaussian vector

$$(X_{f_1}, \cdots, X_{f_L})$$

with known mean function $E[X_f]$ and covariance function $Cov(X_f, X_g)$ that can be calculated from contour integrals involving parameters $\underline{m}(z)$ and H, where $\underline{m}(z)$ is the companion Stieltjes transform corresponding to the limiting SD of $\underline{S}_n = \frac{1}{n} Z_n^* \Sigma_p Z_n$. If the population has a spiked covariance structure, we know that the limit H and $\underline{m}(z)$ remain the same as the non-spiked case, so the limiting parameters $E[X_f]$ and $Cov(X_f, X_g)$ are also unchanged.

It is remarked that the centering parameter $pF^{y_n, H_n}(f)$ depends on a particular distribution F^{y_n, H_n} which is a finite-horizon proxy for the limiting SD of S_n. The difficulty is that F^{y_n, H_n} has no explicit form; it is indeed *implicitly* defined through $\underline{m}_n(z)$ (the finite counterpart of $\underline{m}(z)$), which solves the equation:

$$z = -\frac{1}{\underline{m}_n} + y_n \int \frac{t}{1 + t\underline{m}_n} dH_n(t) \ . \tag{2.3}$$

This distribution depends on the SD H_n which in turn depends on the spike eigenvalues.

More precisely, the SD H_n of Σ_p is

$$H_n = \frac{p - M}{p}\delta_1 + \frac{1}{p}\sum_{i=1}^{k} n_i \delta_{a_i} \ . \tag{2.4}$$

The term

$$\frac{1}{p}\sum_{i=1}^{k} n_i \delta_{a_i}(t)$$

vanishes when p tends to infinity, so it has no influence when considering limiting spectral distributions. However for the CLT, the term $pF^{y_n,H_n}(f)$ has a p in front, and $\frac{1}{p}\sum_{i=1}^{k} n_i\delta_{a_i}$ times p is of order $O(1)$, thus cannot be neglected.

Our main result is an asymptotic expansion for this centering parameter.

Theorem 1. *(large spikes) Suppose the population has a spiked population structure as stated in* (1.1)*, where the spike eigenvalues $a_i's$ satisfy $|a_i-1|>\sqrt{y}$. For any f analytic on an open domain including the support of the Marčenko-Pastur distribution G_y, we have*

$$\begin{aligned}
&F^{y_n,H_n}(f)\\
=&-\frac{1}{2\pi ip}\oint_{\mathcal{C}_1} f(-\frac{1}{m}+\frac{y_n}{1+m})(\frac{M}{y_n m}-\sum_{i=1}^{k}\frac{n_i a_i^2 m}{(1+a_i m)^2})dm \qquad (2.5)\\
+&\frac{1}{2\pi ip}\oint_{\mathcal{C}_1} f^{'}(-\frac{1}{m}+\frac{y_n}{1+m})\sum_{i=1}^{k}\frac{(1-a_i)n_i}{(1+a_i m)(1+m)}(\frac{1}{m}-\frac{y_n m}{(1+m)^2})dm \qquad (2.6)\\
+&(1-\frac{M}{p})G_{y_n}(f)+\frac{1}{p}\sum_{i=1}^{k} n_i f(a_i+\frac{y_n a_i}{a_i-1})+O(\frac{1}{n^2})\ . \qquad (2.7)
\end{aligned}$$

Here $G_{y_n}(f)$ is the integral of f with respect to the Marčenko-Pastur distribution with index $y_n=p/n$ and $\mathcal{C}_1$ is a contour, when restricted to the real axes, enclosing the interval $[\frac{-1}{1-\sqrt{y}},\frac{-1}{1+\sqrt{y}}]$ (or $[\frac{-1}{1+\sqrt{y}},\frac{1}{y-1})\bigcup(\frac{1}{y-1},\frac{-1}{1-\sqrt{y}}]$) when $0<y<1$ (or $y>1$).

Theorem 2. *(small spikes) When the spike eigenvalues $a_i's$ satisfy $|a_i-1|<\sqrt{y}$, we have*

$$\begin{aligned}
&F^{y_n,H_n}(f)\\
=&-\frac{1}{2\pi ip}\oint_{\mathcal{C}_1} f(-\frac{1}{m}+\frac{y_n}{1+m})(\frac{M}{y_n m}-\sum_{i=1}^{k}\frac{n_i a_i^2 m}{(1+a_i m)^2})dm\\
+&\frac{1}{2\pi ip}\oint_{\mathcal{C}_1} f^{'}(-\frac{1}{m}+\frac{y_n}{1+m})\sum_{i=1}^{k}\frac{(1-a_i)n_i}{(1+a_i m)(1+m)}(\frac{1}{m}-\frac{y_n m}{(1+m)^2})dm\\
+&(1-\frac{M}{p})G_{y_n}(f)+O(\frac{1}{n^2})\ ,
\end{aligned}$$

here the $G_{y_n}(f)$ and the contour $\mathcal{C}_1$ are the same as claimed in Theorem 1 above.

Proof. (proof of Theorem 1) Recall that $G_{y_n}(f)=\int f(x)dG_{y_n}(x)$ when no spike exists, where G_{y_n} is the $M-P$ distribution with index y_n. And by the Cauchy integral formula, it can be expressed as $-\frac{1}{2\pi i}\oint f(z)m(z)dz$, where the integral

contour is chosen to enclose the support of G_{y_n} and it's limit G_y. Besides, $\underline{m}$(the companion Stieltjes transform of m) satisfies the equation:

$$z = -\frac{1}{\underline{m}} + \frac{y_n}{1+\underline{m}} \ ,$$

taking derivatives on both sides with respect to z, we get:

$$dz = (\frac{1}{\underline{m}^2} - \frac{y_n}{(1+\underline{m})^2})d\underline{m} \ .$$

Combine these up and use the relationship between m and $\underline{m}$, we have:

$$\begin{aligned} G_{y_n}(f) &= -\frac{n}{p}\frac{1}{2\pi i}\oint f(z)\underline{m}(z)dz \\ &= -\frac{n}{p}\frac{1}{2\pi i}\oint_{\mathcal{C}_1} f(-\frac{1}{\underline{m}} + \frac{y_n}{1+\underline{m}})\underline{m}(z)(\frac{1}{\underline{m}^2} - \frac{y_n}{(1+\underline{m})^2})d\underline{m} \ , \end{aligned}$$

where the second equality is due to the change of variable from z to $\underline{m}$. As a result, the integral region, which encloses the support of G_{y_n} and G_y, is transformed into another region, denoted as $\mathcal{C}_1$. In all the following, we write m to represent $\underline{m}$ for short and it should be noticed that m in this paper is not the Stieltjes transform as we usually denote.

When the spiked structure (1.1) exists, by equation (2.3), this time the Stieltjes transform $m = m_n$ of F^{S_n} satisfies

$$\begin{aligned} z &= -\frac{1}{m} + \frac{p-M}{p}\frac{y_n}{1+m} + \frac{y_n}{p}\sum_{i=1}^{k}\frac{a_i n_i}{1+a_i m} \ , \\ dz &= \left(\frac{1}{m^2} - \frac{p-M}{p}\frac{y_n}{(1+m)^2} - \frac{y_n}{p}\sum_{i=1}^{k}\frac{a_i^2 n_i}{(1+a_i m)^2}\right)dm \ . \end{aligned}$$

Repeating the same computation as before, we get:

$$\begin{aligned} & F^{y_n, H_n}(f) \\ = \ & -\frac{n}{p}\frac{1}{2\pi i}\oint_{\mathcal{C}} f(z)m\left(\frac{1}{m^2} - \frac{p-M}{p}\frac{y_n}{(1+m)^2} - \frac{y_n}{p}\sum_{i=1}^{k}\frac{a_i^2 n_i}{(1+a_i m)^2}\right)dm \\ = \ & -\frac{n}{p}\frac{1}{2\pi i}\oint_{\mathcal{C}} f\Big(-\frac{1}{m} + \frac{y_n}{1+m} - \frac{y_n}{p}\sum_{i=1}^{k}\frac{(1-a_i)n_i}{(1+m)(1+a_i m)}\Big)m \\ & \times\left(\frac{1}{m^2} - \frac{y_n}{(1+m)^2} + \frac{y_n}{p}\sum_{i=1}^{k} n_i\left[\frac{1}{(1+m)^2} - \frac{a_i^2}{(1+a_i m)^2}\right]\right)dm \ . \end{aligned}$$

The term

$$\frac{y_n}{p}\sum_{i=1}^{k}\frac{(1-a_i)n_i}{(1+m)(1+a_i m)}$$

is of order $O(\frac{1}{n})$, so we can take the Taylor expansion of f around the value of $-\frac{1}{m}+\frac{y_n}{1+m}$, and the term

$$\frac{y_n}{p}\sum_{i=1}^{k} n_i\Big[\frac{1}{(1+m)^2}-\frac{a_i^2}{(1+a_im)^2}\Big]$$

is also of order $O(\frac{1}{n})$, this gives rise to:

$$\begin{aligned}
F^{y_n,H_n}(f) &= -\frac{n}{p}\frac{1}{2\pi i}\oint_{\mathcal{C}} f(-\frac{1}{m}+\frac{y_n}{1+m})(\frac{1}{m}-\frac{y_nm}{(1+m)^2})dm\\
&-\frac{n}{p}\frac{1}{2\pi i}\oint_{\mathcal{C}} f(-\frac{1}{m}+\frac{y_n}{1+m})\frac{y_n}{p}\sum_{i=1}^{k} n_i\Big[\frac{1}{(1+m)^2}-\frac{a_i^2}{(1+a_im)^2}\Big]mdm\\
&+\frac{n}{p}\frac{1}{2\pi i}\oint_{\mathcal{C}} f^{'}(-\frac{1}{m}+\frac{y_n}{1+m})\frac{y_n}{p}\sum_{i=1}^{k}\frac{(1-a_i)n_i}{(1+m)(1+a_im)}(\frac{1}{m}-\frac{y_nm}{(1+m)^2})dm\\
&+O(\frac{1}{n^2})\ . \qquad (2.8)
\end{aligned}$$

First we consider these $\mathcal{C}_{a_i}(i=1,\cdots,k)$. From the formula (2.8), we see that only the last three terms contribute to the integral on the contour $\mathcal{C}_{a_i}(i=1,\cdots,k)$ (because for the first term, the only poles: $m=0$ and $m=-1$ are not in the contour C_{a_i}):

$$\begin{aligned}
&-\frac{n}{p}\frac{1}{2\pi i}\oint_{\mathcal{C}_{a_i}} f(-\frac{1}{m}+\frac{y_n}{1+m})\frac{y_n}{p}\sum_{i=1}^{k} n_i\Big[\frac{1}{(1+m)^2}-\frac{a_i^2}{(1+a_im)^2}\Big]mdm\\
=\ &\frac{n}{p}\frac{1}{2\pi in}\oint_{\mathcal{C}_{a_i}} f(-\frac{1}{m}+\frac{y_n}{1+m})\sum_{i=1}^{k}\frac{a_i^2n_im}{(1+a_im)^2}dm\\
=\ &\frac{1}{2\pi ip}\oint_{\mathcal{C}_{a_i}}\frac{f(-\frac{1}{m}+\frac{y_n}{1+m})mn_i}{(m+\frac{1}{a_i})^2}dm\\
=\ &\frac{n_i}{p}\Big[f(\phi(a_i))-f^{'}(\phi(a_i))\Big(a_i-\frac{y_na_i}{(a_i-1)^2}\Big)\Big]\ ,
\end{aligned}$$

$$\begin{aligned}
&\frac{n}{p}\frac{1}{2\pi i}\oint_{\mathcal{C}_{a_i}} f^{'}(-\frac{1}{m}+\frac{y_n}{1+m})\frac{y_n}{p}\sum_{i=1}^{k}\frac{(1-a_i)n_i}{(1+m)(1+a_im)}(\frac{1}{m}-\frac{y_nm}{(1+m)^2})dm\\
=\ &\frac{-1}{2\pi ip}\oint_{\mathcal{C}_{a_i}} f^{'}(-\frac{1}{m}+\frac{y_n}{1+m})\frac{n_i(1-a_i)}{(m+\frac{1}{a_i})a_i(m+1)}(\frac{1}{m}-\frac{y_nm}{(1+m)^2})dm\\
=\ &\frac{1}{p}n_if^{'}(\phi(a_i))\Big(a_i-\frac{y_na_i}{(a_i-1)^2}\Big).
\end{aligned}$$

So, combining these two terms, we get the influence of the spiked part, that is, the integral on the contours $\bigcup_{i=1,\cdots k}\mathcal{C}_{a_i}$:

$$\frac{1}{p}\sum_{i=1}^{k} n_if(\phi(a_i))\ . \qquad (2.9)$$

So in the remaining part, we only need to consider the integral along the contour $\mathcal{C}_1$. Consider the second term of (2.8) with the contour being $\mathcal{C}_1$:

$$\begin{aligned}
& -\frac{n}{p}\frac{1}{2\pi i}\oint_{\mathcal{C}_1} f(-\frac{1}{m}+\frac{y_n}{1+m})\frac{y_n}{p}\sum_{i=1}^{k} n_i\left[\frac{1}{(1+m)^2}-\frac{a_i^2}{(1+a_i m)^2}\right]m dm \\
= & -\frac{1}{2\pi i p}\oint_{\mathcal{C}_1} f(-\frac{1}{m}+\frac{y_n}{1+m})\left[\frac{1}{y_n}(\frac{Mmy_n}{(1+m)^2}-\frac{M}{m})+\frac{1}{y_n}\frac{M}{m}-\sum_{i=1}^{k}\frac{n_i a_i^2 m}{(1+a_i m)^2}\right]dm \\
= & -\frac{M}{p}G_{y_n}(f)-\frac{1}{2\pi i p}\oint_{\mathcal{C}_1} f(-\frac{1}{m}+\frac{y_n}{1+m})\left(\frac{M}{my_n}-\sum_{i=1}^{k}\frac{n_i a_i^2 m}{(1+a_i m)^2}\right)dm \qquad (2.10)
\end{aligned}$$

Combining Equations (2.8), (2.9) and (2.10), we get:

$$\begin{aligned}
F^{y_n,H_n}(f) = & -\frac{1}{2\pi i p}\oint_{\mathcal{C}_1} f(-\frac{1}{m}+\frac{y_n}{1+m})(\frac{M}{my_n}-\sum_{i=1}^{k}\frac{n_i a_i^2 m}{(1+a_i m)^2})dm \\
& +\frac{1}{2\pi i p}\oint_{\mathcal{C}_1} f'(-\frac{1}{m}+\frac{y_n}{1+m})\sum_{i=1}^{k}\frac{(1-a_i)n_i}{(1+m)(1+a_i m)}(\frac{1}{m}-\frac{y_n m}{(1+m)^2})dm \\
& +(1-\frac{M}{p})G_{y_n}(f)+\sum_{i=1}^{k}\frac{n_i}{p}f(\phi(a_i))+O(\frac{1}{n^2})\ .
\end{aligned}$$

The proof of the theorem is complete. □

Proof. (proof of Theorem 2) When the $a_i's$ satisfy $|a_i-1|<\sqrt{y}$, from Baik and Silverstein (2006), we know that the extreme eigenvalues of S_n converge to the endpoints of the support of the M-P distribution, i.e a_y or b_y. So the term $\frac{1}{p}\sum_{i=1}^{k} n_i f(a_i+\frac{y_n a_i}{a_i-1})$ in Equation (2.7), which corresponds to the isolated spiked part is absent, and the contour $\mathcal{C}$ is now exactly the contour $\mathcal{C}_1$ which contains only the interval $[\frac{-1}{1+\sqrt{y}},\frac{1}{y-1})\bigcup(\frac{1}{y-1},\frac{-1}{1-\sqrt{y}}]$ when $y>1$ and $[\frac{-1}{1-\sqrt{y}},\frac{-1}{1+\sqrt{y}}]$ when $0<y<1$. Taking Equation (2.10) into consideration leads to the result. □

3. About the contour $\mathcal{C}$ and $\mathcal{C}_1$ appearing in the main theorems and proofs

As we can see in the derivations of Theorem 1 and 2, one of the most technical points concerns contour integrals on $\mathcal{C}$ and $\mathcal{C}_1$. In this section, we provide more details on these contour integrals.

When no spike exists, we have the equation: $z=-\frac{1}{m}+\frac{y_n}{1+m}$. So the transform: $z\longrightarrow m(z)$ maps a region in the complex plan that includes the support of G_{y_n} and G_y to a region of m, which we denote as $\mathcal{C}_1$. This mapping is not easy to visualize in the complex plan; we can consider its restriction to the real domains, i.e. for real z and m, as illustrated in Figure 1, with left panel corresponds to $0<y<1$ and right panel $y>1$. Because the two situations lead to different

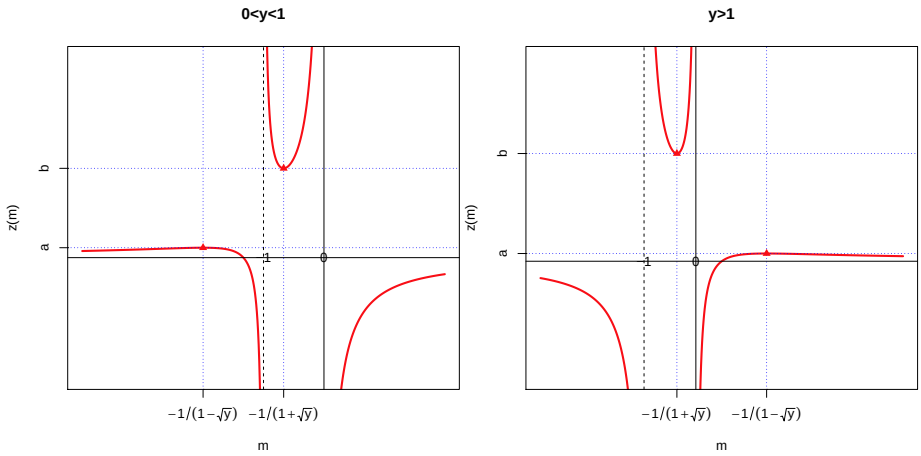

Figure 1: The graph of the transform $z(m) = -\frac{1}{m} + \frac{y}{1+m}$.

appearance of the graph, we should consider them separately. Both two panels have two local extrema located at $(-\frac{1}{1-\sqrt{y}}, a)$ and $(-\frac{1}{1+\sqrt{y}}, b)$, where a and b are the endpoints of the support of the M-P distribution.

The $z's$ such that $z'(m) > 0$ are not in the support of F^{S_n}(or G_{y_n}), refer to Silverstein and Choi (1995) for further reference. First consider the case of $0 < y < 1$: from the left penal of Figure 1, we see that the $z's$ such that $z'(m) > 0$ corresponds to $\{z > b\} \bigcup \{0 < z < a\} \bigcup \{z < 0\}$. Thus, the support of F^{S_n} consists of $\{a \leq z \leq b\} \bigcup \{z = 0\}$, and changing to the variable m, it is equivalent to $m \in [\frac{-1}{1-\sqrt{y}}, \frac{-1}{1+\sqrt{y}}]$. From the figure, we see that the point $\{m = \frac{1}{y-1}\}$ (equivalent to $\{z = 0\}$) is contained in this interval. As a result, restrict the contour $\mathcal{C}_1$ to the real axes, it encloses the interval $[\frac{-1}{1-\sqrt{y}}, \frac{-1}{1+\sqrt{y}}]$ when $0 < y < 1$. Besides, $\{m = -1\}$ is the pole. For the case of $y > 1$: from the right panel of Figure 1, $\{z'(m) > 0\}$ corresponds to $\{z > b\} \bigcup \{-\infty < z < a\}$, thus, the support of F^{S_n}(or G_{y_n}) consists of $\{a \leq z \leq b\}$, which is equivalent to $m \in [\frac{-1}{1+\sqrt{y}}, \frac{-1}{1-\sqrt{y}}]$. Since in this situation, $\{z = 0\} \notin \{a \leq z \leq b\}$, so the value of $m = \frac{1}{y-1}$ (equivalent to $z = 0$) should be excluded from $[\frac{-1}{1+\sqrt{y}}, \frac{-1}{1-\sqrt{y}}]$. For this reason, the contour $\mathcal{C}_1$ encloses $[\frac{-1}{1+\sqrt{y}}, \frac{1}{y-1}) \bigcup (\frac{1}{y-1}, \frac{-1}{1-\sqrt{y}}]$ when $y > 1$. Besides, this time the pole is $\{m = 0\}$.

Furthermore, suppose there exists a spiked structure as (1.1), so the equation becomes

$$z = -\frac{1}{m} + \frac{p-M}{p} \cdot \frac{y_n}{1+m} + \frac{y_n}{p} \sum_{i=1}^{k} \frac{a_i n_i}{1+a_i m} \ ,$$

the transform: $z \longrightarrow m(z)$ maps a region of z (containing $supp(F^{S_n})$ and $supp(F^S)$) into $\mathcal{C}$. This time, whether the spikes are large or small leads to

different situations. When $|a_i - 1| > \sqrt{y}$ (large spikes), the spike eigenvalues a_i's are away from the critical values $1 \pm \sqrt{y}$, and it has been proven in Baik and Silverstein (2006) that the support of F^{S_n} is made of $k+1$ absolutely continuous components exactly: a bulk part clustering around the support of the M-P distribution G_{y_n} and k isolated small components with support $[a_{a_i}, b_{a_i}]$ clustering around the value $a_i + ya_i/(a_i - 1) \triangleq \phi(a_i)$, $i = 1, \ldots, k$, respectively. Moreover, due to the exact separation property, $F^{y_n, H_n}[a_{a_i}, b_{a_i}] = n_i/p$, exactly for large p. For this reason, after mapping the support of F^{S_n} to $\mathcal{C}$, it also consists of $k+1$ parts: $\mathcal{C}_1$ plus $C_{a_i}(i = 1, \cdots, k)$, with each of these contour non-overlapping. And $\mathcal{C}_{a_i}$ is the contour that encloses the point of $\{m = -\frac{1}{a_i}\}$ (because the k isolated parts surround around the value of $z = \phi(a_i)(i = 1, \cdots, k)$ and mapping these values to m is exactly $m = -\frac{1}{a_i}(i = 1, \cdots, k)$). When $|a_i - 1| < \sqrt{y}$ (small spikes), Baik and Silverstein (2006) shows that the extreme eigenvalues of S_n cannot be separated from the support of the M-P distribution (they stick to the endpoints of the support: a or b). So the contour $\mathcal{C}$ is now just $\mathcal{C}_1$(without the other k components $\mathcal{C}_{a_i}(i = 1, \cdots, k)$). Moreover, since

$$z = -\frac{1}{m} + \frac{p - M}{p} \cdot \frac{y_n}{1 + m} + \frac{y_n}{p} \sum_{i=1}^{k} \frac{a_i n_i}{1 + a_i m} ,$$

we see that $m = -\frac{1}{a_i}$ is the pole. Due to $|a_i - 1| < \sqrt{y}$, we have always $-\frac{1}{a_i} \in [-\frac{1}{1-\sqrt{y}}, -\frac{1}{1+\sqrt{y}}]$ or $-\frac{1}{a_i} \in [-\frac{1}{1+\sqrt{y}}, \frac{1}{y-1}) \bigcup (\frac{1}{y-1}, -\frac{1}{1-\sqrt{y}}]$ according to whether $0 < y < 1$ or $y > 1$.

We may summarize all these findings in the following Table 1 and 2. Besides, it should be noticed that for a given function f, there will be other poles which belongs to f that contribute to the integral if these poles are in the region of $\mathcal{C}_1$.

Table 1

	$0 < y < 1$	$y > 1$
region of $\mathcal{C}_1$	$m \in [-\frac{1}{1-\sqrt{y}}, -\frac{1}{1+\sqrt{y}}]$	$m \in [-\frac{1}{1+\sqrt{y}}, \frac{1}{y-1}) \bigcup (\frac{1}{y-1}, -\frac{1}{1-\sqrt{y}}]$

Table 2

poles in $\mathcal{C}_1$	$0 < y < 1$	$y > 1$
large spikes	$m = -1$	$m = 0$
small spikes	$m = -1, m = -\frac{1}{a_i}(i = 1, \cdots, k)$	$m = 0, m = -\frac{1}{a_i}(i = 1, \cdots, k)$

Last, we compare the results of Theorem 1 and Theorem 2. Seemly, Theorem 1 has one term more than Theorem 2 (the term $\frac{1}{p}\sum_{i=1}^{k} n_i f(a_i + \frac{y_n a_i}{a_i - 1})$ is missing in Theorem 2). But according to Table 2, we know that when doing contour integral of small spikes, the terms that correspond to the contribution of poles $m = -\frac{1}{a_i}(i = 1, \cdots, k)$ in $\mathcal{C}_1$ should be added. And it is exactly the missing term $\frac{1}{p}\sum_{i=1}^{k} n_i f(a_i + \frac{y_n a_i}{a_i - 1})$ appearing in Theorem 1 if we do the calculation. So, Theorem 1 and Theorem 2 lead to the same result in spite of different expressions. Above all, we have the knowledge that whether the spikes are large or small has no influence on the centering parameter $F^{y_n, H_n}(f)$.

4. An application to the test of presence of spike eigenvalues

In Bai et al. (2009), a corrected likelihood ratio statistic $\tilde{L}^*$ is proposed to test the hypothesis

$$H_0 : \Sigma = I_p \quad vs. \quad H_1 : \Sigma \neq I_p \ .$$

They prove that under H_0,

$$\tilde{L}^* - pG_{y_n}(g) \Rightarrow N(m(g), v(g)) \ ,$$

where

$$\begin{aligned}
&\tilde{L}^* = trS_n - \log|S_n| - p \ , \\
&G_{y_n}(g) = 1 - \frac{y_n - 1}{y_n}\log(1 - y_n) \ , \\
&m(g) = -\frac{\log(1-y)}{2} \ , \\
&v(g) = -2\log(1-y) - 2y.
\end{aligned}$$

At a significance level α (usually 0.05), the test will reject H_0 when $\tilde{L}^* - pG_{y_n}(g) > m(g) + \Phi^{-1}(1-\alpha)\sqrt{v(g)}$ where Φ is the standard normal cumulative distribution function.

However, the power function of this test remains unknown because the distribution of $\tilde{L}^*$ under the general alternative hypothesis H_1 is ill-defined. Let's consider this general test as a way to test the null hypothese H_0 above against an alternative hypothesis of the form:

$$H_1^* : \Sigma \text{ has the spiked structure (1.1).}$$

In other words, we want to test the absence against the presence of possible spike eigenvalues in the population covariance matrix. The general asymptotic expansion in Theorem 1 helps to find the power function of the test when the spike eigenvalues are large. However, it will lead to the same result if we use Theorem 2.

More precisely, under the alternative H_1^* and for $f(x) = x - \log x - 1$ used in the statistic $\tilde{L}^*$, the centering term $F^{y_n, H_n}(f)$ can be found to be

$$1 + \frac{1}{p}\sum_{i=1}^{k} n_i a_i - \frac{M}{p} - \frac{1}{p}\sum_{i=1}^{k} n_i \log a_i - (1 - \frac{1}{y_n})\log(1 - y_n) + O(\frac{1}{n^2}) \ ,$$

thanks to the following formula

$$F^{y_n, H_n}(x) = 1 + \frac{1}{p}\sum_{i=1}^{k} n_i a_i - \frac{M}{p} + O(\frac{1}{n^2}) \tag{4.11}$$

and

$$F^{y_n,H_n}(\log x) = \frac{1}{p}\sum_{i=1}^{k} n_i \log a_i - 1 + (1-\frac{1}{y_n})\log(1-y_n) + O(\frac{1}{n^2}) \ . \quad (4.12)$$

The details of derivation of these formula are given in the Appendix.

Therefore we have obtained that under H_1^*,

$$\tilde{L}^* - pF^{y_n,H_n}(f) \Rightarrow N(m(g), v(g)) \ .$$

It follows that the asymptotic power function of the test is

$$\beta(\alpha) = 1 - \Phi\left(\Phi^{-1}(1-\alpha) - \frac{\sum_{i=1}^{k} n_i a_i - M - \sum_{i=1}^{k} n_i \log a_i}{\sqrt{-2\log(1-y) - 2y}}\right) .$$

In the particular case where the spiked model has only one simple spike, i.e. $k = 1$, $n_1 = 1$, the above power function becomes

$$\beta(\alpha) = 1 - \Phi\left(\Phi^{-1}(1-\alpha) - \frac{a_1 - 1 - \log a_1}{\sqrt{-2\log(1-y) - 2y}}\right) \ ,$$

which is exactly the formula (5.6) found in Onatski et al. (2011). Note that these authors have found this formula using a sophisticated tools of asymptotic contiguity and Le Cam's first and third lemmas. Our derivation is in a sense much more direct.

Appendix A: Additional proofs of (4.11) and (4.12)

Recall that the LRT test works only when $0 < y < 1$, so the contour $\mathcal{C}_1$ encloses the interval $[\frac{-1}{1-\sqrt{y}}, \frac{-1}{1+\sqrt{y}}]$ on the real axes. And poles of $\{m = -1\}$ and $\{m = \frac{1}{y_n - 1}\}$ (pole of the function $logz$) are included in this interval. In all the following, we only consider large spikes and do calculations according to Theorem 1.

A.1. Proof of (4.11)

We have

$$\begin{aligned}
(2.5) &= -\frac{1}{2\pi i p}\oint_{\mathcal{C}_1}(-\frac{1}{m} + \frac{y_n}{1+m})(\frac{M}{y_n m} - \sum_{i=1}^{k}\frac{n_i a_i^2 m}{(1+a_i m)^2})dm \\
&= -\frac{1}{2\pi i p}\oint_{\mathcal{C}_1}\left(\frac{M}{m(m+1)} - y_n\sum_{i=1}^{k}\frac{n_i a_i^2 m}{(1+a_i m)^2(1+m)}\right)dm \\
&= \frac{M}{p} - \frac{y_n}{p}\sum_{i=1}^{k}\frac{n_i a_i^2}{(1-a_i)^2} \ , \qquad (A.13)
\end{aligned}$$

$$
\begin{aligned}
(2.6) &= \frac{1}{2\pi ip}\oint_{\mathcal{C}_1}\sum_{i=1}^{k}\frac{(1-a_i)n_i}{(1+a_im)(1+m)}\Big(\frac{1}{m}-\frac{y_nm}{(1+m)^2}\Big)dm \\
&= \frac{1}{p}\sum_{i=1}^{k}\Big[-n_i-\frac{1}{2}(1-a_i)n_iy_n\frac{\partial}{\partial m^2}\Big(\frac{m}{1+a_im}\Big)^2\Big|_{m=-1}\Big] \\
&= \frac{1}{p}\sum_{i=1}^{k}\Big[-n_i+\frac{a_in_iy_n}{(1-a_i)^2}\Big]\ , \qquad (A.14)
\end{aligned}
$$

$$
(2.7) = 1-\frac{M}{p}+\frac{1}{p}\sum_{i=1}^{k}n_i\Big(a_i+\frac{y_na_i}{a_i-1}\Big)+O(\frac{1}{n^2})\ . \qquad (A.15)
$$

Combine (A.13), (A.14) and (A.15), we get:

$$
F^{y_n,H_n}(x)=1+\frac{1}{p}\sum_{i=1}^{k}n_ia_i-\frac{M}{p}+O(\frac{1}{n^2})\ .
$$

A.2. Proof of (4.12)

We have

$$
\begin{aligned}
(2.5) &= \frac{-1}{2\pi ipy_n}\oint_{\mathcal{C}_1}\frac{\log(\frac{y_n-1}{m})+\log(\frac{m-\frac{1}{y_n-1}}{m+1})}{m}(M-\sum_{i=1}^{k}\frac{n_ia_i^2y_nm^2}{(1+a_im)^2})dm \\
&= \frac{-M}{2\pi ipy_n}\oint_{\mathcal{C}_1}\frac{\log(\frac{m-\frac{1}{y_n-1}}{m+1})}{m}dm+\frac{1}{2\pi ipy_n}\oint_{\mathcal{C}_1}\log(\frac{m-\frac{1}{y_n-1}}{m+1})\sum_{i=1}^{k}\frac{n_ia_i^2y_nm}{(1+a_im)^2}dm \\
&\triangleq A+B\ . \qquad (A.16)
\end{aligned}
$$

$$
\begin{aligned}
A &= \frac{-M}{2\pi ipy_n}\oint_{\mathcal{C}_1}\log(\frac{m-\frac{1}{y_n-1}}{m+1})\cdot d\log m \\
&= \frac{M}{2\pi ipy_n}\oint_{\mathcal{C}_1}\log m\cdot d\log(\frac{m-\frac{1}{y_n-1}}{m+1}) \\
&= \frac{M}{2\pi ipy_n}\cdot\frac{y_n}{y_n-1}\oint_{\mathcal{C}_1}\frac{\log m}{(m+1)(m-\frac{1}{y_n-1})}dm \\
&= -\frac{M}{py_n}\log(1-y_n)\ , \qquad (A.17)
\end{aligned}
$$

$$\begin{aligned} B &= \frac{1}{2\pi ip}\oint_{\mathcal{C}_1}\log(\frac{m-\frac{1}{y_n-1}}{m+1})\sum_{i=1}^{k}\frac{n_i a_i^2 m}{(1+a_i m)^2}dm \\ &= \frac{1}{2\pi ip}\sum_{i=1}^{k}\oint_{\mathcal{C}_1}\log(\frac{m-\frac{1}{y_n-1}}{m+1})n_i a_i(\frac{1}{1+a_i m}-\frac{1}{(1+a_i m)^2})dm \\ &\triangleq C-D\ , \end{aligned} \tag{A.18}$$

where

$$\begin{aligned} C &= \frac{1}{2\pi ip}\sum_{i=1}^{k}\oint_{\mathcal{C}_1}\log(\frac{m-\frac{1}{y_n-1}}{m+1})\frac{n_i a_i}{1+a_i m}dm \\ &= \frac{1}{2\pi ip}\sum_{i=1}^{k}\oint_{\mathcal{C}_1} n_i\log(\frac{m-\frac{1}{y_n-1}}{m+1})\cdot d\log(1+a_i m) \\ &= \frac{-1}{2\pi ip}\sum_{i=1}^{k}\oint_{\mathcal{C}_1} n_i\log(1+a_i m)\cdot d\log(\frac{m-\frac{1}{y_n-1}}{m+1}) \\ &= \frac{-1}{2\pi ip}\cdot\frac{y_n}{y_n-1}\sum_{i=1}^{k}\oint_{\mathcal{C}_1}\frac{n_i\log(1+a_i m)}{(m+1)(m-\frac{1}{y_n-1})}dm \\ &= \frac{1}{p}\sum_{i=1}^{k}n_i\log(1-a_i)-\frac{1}{p}\sum_{i=1}^{k}n_i\log(1+\frac{a_i}{y_n-1})\ , \end{aligned} \tag{A.19}$$

and

$$\begin{aligned} D &= \frac{1}{2\pi ip}\sum_{i=1}^{k}\oint_{\mathcal{C}_1}\log(\frac{m-\frac{1}{y_n-1}}{m+1})\frac{n_i a_i}{(1+a_i m)^2}dm \\ &= \frac{1}{2\pi ip}\sum_{i=1}^{k}\oint_{\mathcal{C}_1}\frac{n_i}{1+a_i m}\cdot d\log(\frac{m-\frac{1}{y_n-1}}{m+1}) \\ &= \frac{y_n}{2\pi ip(y_n-1)}\sum_{i=1}^{k}\oint_{\mathcal{C}_1}\frac{n_i}{(1+a_i m)(m-\frac{1}{y_n-1})(m+1)}dm \\ &= \frac{1}{p}\sum_{i=1}^{k}(\frac{n_i}{1+\frac{a_i}{y_n-1}}-\frac{n_i}{1-a_i})\ . \end{aligned} \tag{A.20}$$

Combine (A.16), (A.17), (A.18), (A.19) and (A.20), we get:

$$\begin{aligned} (2.5) &= -\frac{M}{py_n}\log(1-y_n)+\frac{1}{p}\sum_{i=1}^{k}n_i\log(1-a_i)-\frac{1}{p}\sum_{i=1}^{k}n_i\log(1+\frac{a_i}{y_n-1}) \\ &\quad -\frac{1}{p}\sum_{i=1}^{k}\frac{n_i}{1+\frac{a_i}{y_n-1}}+\frac{1}{p}\sum_{i=1}^{k}\frac{n_i}{1-a_i}\ . \end{aligned} \tag{A.21}$$

Then, we consider the part (2.6) in the general formula:

$$\begin{aligned}(2.6) &= -\frac{1}{2\pi ip}\oint_{\mathcal{C}_1} f^{'}(-\frac{1}{m}+\frac{y_n}{1+m})\sum_{i=1}^{k}(\frac{n_i a_i}{1+a_i m}-\frac{n_i}{1+m})(\frac{1}{m}-\frac{y_n m}{(1+m)^2})dm \\ &= -\frac{1}{2\pi ip}\sum_{i=1}^{k} n_i \oint_{\mathcal{C}_1}\frac{m(m+1)}{y_n m-m-1}(\frac{a_i}{1+a_i m}-\frac{1}{1+m})(\frac{1}{m}-\frac{y_n m}{(1+m)^2})dm \\ &\triangleq \frac{-1}{2\pi ip(y_n-1)}\sum_{i=1}^{k} n_i(E-F-G+H)\ ,\end{aligned}$$

where

$$\begin{aligned}E &= \oint_{\mathcal{C}_1}\frac{a_i(m+1)}{(1+a_i m)(m-\frac{1}{y_n-1})} = 2\pi i\frac{y_n a_i}{y_n+a_i-1}\ , \\ F &= \oint_{\mathcal{C}_1}\frac{a_i y_n m^2}{(m+1)(1+a_i m)(m-\frac{1}{y_n-1})} = 2\pi i(\frac{a_i(y_n-1)}{a_i-1}+\frac{a_i}{y_n+a_i-1})\ , \\ G &= \oint_{\mathcal{C}_1}\frac{1}{m-\frac{1}{y_n-1}} = 2\pi i\ , \\ H &= \oint_{\mathcal{C}_1}\frac{y_n m^2}{(m+1)^2(m-\frac{1}{y_n-1})}dm = 2\pi i y_n\ .\end{aligned}$$

Collecting these four terms, we have:

$$(2.6) = \frac{1}{p}\sum_{i=1}^{k}(\frac{1}{a_i-1}-\frac{a_i}{y_n+a_i-1})n_i\ . \tag{A.22}$$

Finally, using the known result that $G_{y_n}(\log x) = (1-\frac{1}{y_n})\log(1-y_n)-1$, which has been calculated in Bai and Silverstein (2004), and combine (A.21), (A.22) and (2.7), we get

$$F^{y_n,H_n}(\log x) = \frac{1}{p}\sum_{i=1}^{k} n_i\log a_i - 1 + (1-\frac{1}{y_n})\log(1-y_n) + O(\frac{1}{n^2})\ .$$

References

Bai, Z.D. and Silverstein, J.W. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Ann. Probab.*, **32**, 553–605.

Bai, Z.D. and Yao, J.F. (2008). CLT for eigenvalues in a spiked population model. *Ann. Inst. Henri Poincaré Probab. Stat.*, **44**(3), 447–474.

Bai, Z. D., Jiang, D., Yao, J. and Zheng, S. (2009). Corrections to LRT on large dimensional covariance matrix by RMT. *Ann. Statist.* **37**, 3822-3840

Bai, Z.D. and Silverstein, J.W. (2010). *Spectral Analysis of Large Dimensional Random Matrices* (2nd edition). Springer, 20.

Bai, Z.D. and Yao, J.F. (2012). On sample eigenvalues in a generalized spiked population model. *J. Multivariate Anal.*, **106**, 167–177.

Baik, J., Ben Arous, G., and Péché, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.*, **33**(5), 1643–1697.

Baik, J. and Silverstein, J.W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate Anal.*, **97**, 1382–1408.

Benaych-Georges, F., Guionnet, A. and Maida, M. (2011). Fluctuations of the extreme eigenvalues of finite rank deformations of random matrices. *Electron. J. Probab.*, **16**, 1621–1662.

Benaych-Georges, F. and Nadakuditi, R.R. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Adv. Math.*, **227**(2), 494–521.

Johnstone, I.M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, **29**(2), 295–327.

Kritchman, S. and Nadler, B. (2008) Determining the number of components in a factor model from limited noisy data. *Chem. Int. Lab. Syst.* **94**, 19-32

Kritchman, S. and Nadler, B. (2009) Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory. *IEEE Trans. Signal Process.* **57**(10), 3930–3941

Marčenko, V.A. and Pastur, L.A. (1967). Distribution of eigenvalues for some sets of random matrices. *Math. USSR-Sb*, **1**, 457–483

Nadakuditi, R. R. and Silverstein, J.(2010) Fundamental Limit of Sample Generalized Eigenvalue Based Detection of Signals in Noise Using Relatively Few Signal-Bearing and Noise-Only Samples. *IEEE J. Sel. Topics Signal Processing.* **4**(3), 468-480

Onatski, A. (2009) Testing hypotheses about the number of factors in large factor models. *Econometrica* **77** (5), 1447-1479.

Onatski, A. (2010) Determining the number of factors from empirical distribution of eigenvalues. *Review of Economics and Statistics* **92** (4), 1004-1016.

Onatski, A., Moreira, M.J. and Hallin, M. (2011). Asymptotic power of sphericity tests for high-dimensional data. *Preprint*, available at `arXiv:1210.5663v1`.

Onatski, A. (2012) Asymptotics of the principal components estimator of large factor models with weakly influential factors. *J. Econometrics*, **168**, 244-258

Pan, G.M. and Zhou, W. (2008). Central limit theorem for signal-to-interference ratio of reduced rank linear receiver. *Ann. Appl. Probab.*, **18**, 1232-1270.

Passemier, D. and Yao, J.F. (2012). On determining the number of spikes in a high-dimensional spiked population model. *Random Matrix: Theory and Applciations* **1**, 1150002

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica.*, **17**, 1617–1642.

Silverstein, J.W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *J. Multivariate Anal.*, 55 (2):331–339.

Silverstein, J.W. and Choi, S.I. (1995). Analysis of the limiting spectral distri-

bution of large dimensional random matrices. *J. Multivariate Anal.*, 54(2): 295–309.